NoSQL Research Project

Zip Code WIlmington
Data Cohort 2.2

Presented by Keerthi Balla & Tom Lafferty

NoSQL 30,000 foot view

- Stands for “Not Only SQL” or “Non Relational”
- Do away with relationships, schema-less.
  - Does not require fixed table schema nor the concept of joins.
- Every item stands on its own.

➔ Term coined in 1998, although these types of databases have been used since the 1960’s.

Relational vs NoSQL

- Data consistency
- Harder to scale
  - Can scale vertically, not horizontally
- Resource intensive
- At some point, won't be able to handle the load.

- "Eventually consistent"
- Easier to scale
  - Can scale vertically and horizontally
- Partitions are different servers (horizontal)
- Primary key is converted to Hash value
  - Keyspace -> Each partition (server) is assigned a range within the keyspace range
  - How it knows where to store the data, and where to find the data.
  - Can only be retrieved by the primary key
- Partitions are mirrored to other servers to prevent loss from hardware failure
  - May cause data to not be immediately returned.
  - Eventually consistent, it will be returned after milliseconds.

RDBMS –
Relational Database Management System

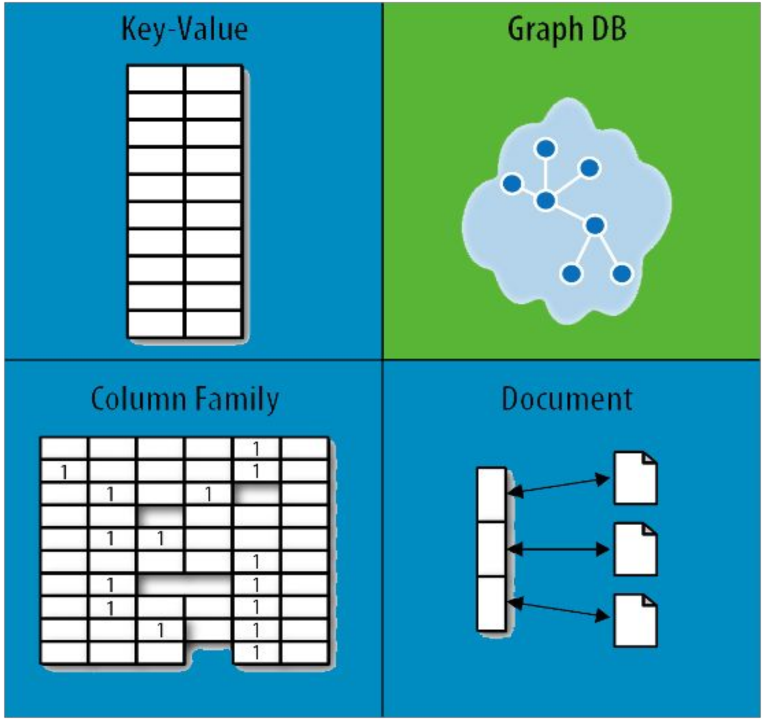
- Relation – a 2D table that has the following features:
  - Name
  - Attributes
  - Tuples
- Issues:
  - Scalability – When the dataset is too big e.g. Big Data
  - Distribution – Not designed to be distributed
  - Horizontal Scaling
  - Different approaches: Master-slave and Sharding

Brief History

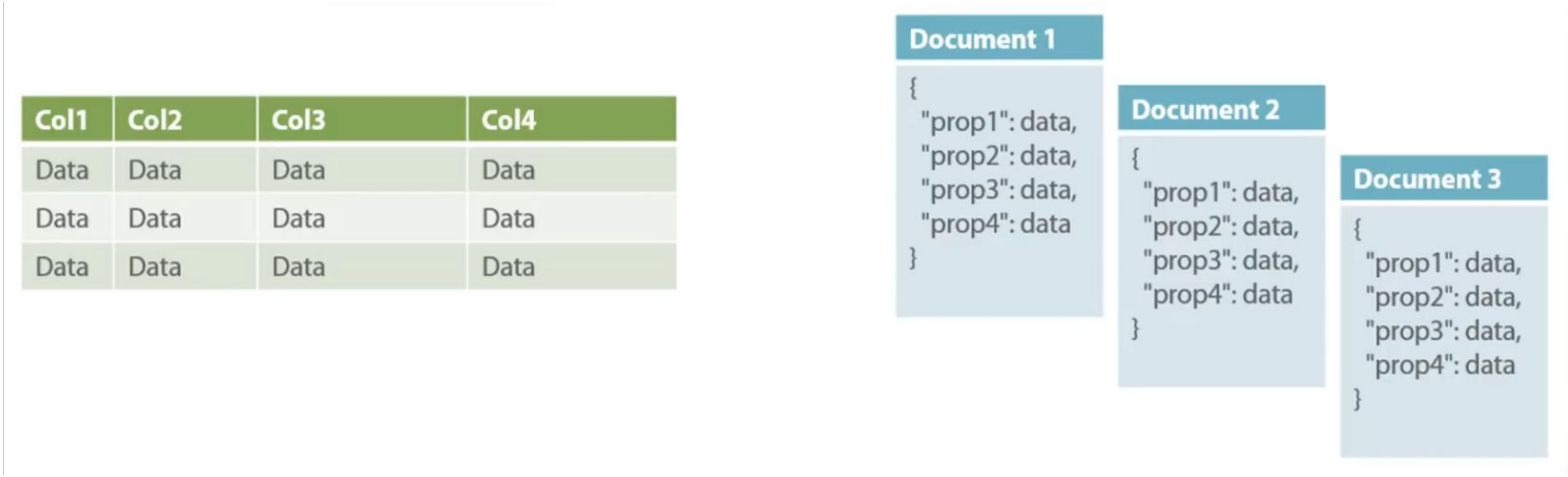
- Non-relational DBMSs are not new
- But NoSQL represents a new incarnation
  - Due to massively scalable Internet applications
  - Based on distributed and parallel computing
- Development
  - Three major papers were the seeds of the NoSQL movement – BigTable(Google), Dynamo (Amazon) [Distributed key-value data store & Eventual Consistency], CAP Theorem
  - Starts with Google
  - First research paper published in 2003
  - Continues also thanks to Lucene's developers/Apache (Hadoop) and Amazon (Dynamo)
  - Then a lot of products and interests came from Facebook, Netflix, Yahoo, eBay, Hulu, IBM, and many more

NoSQL types

- •Document based [MongoDB, CouchDB]
- •Key Value pair based [Redis, Couchbase Server]
- •Column based [HBase]
- •Graph based [Neo4j]



Relational database vs noSQL(document based) database



Col1	Col2	Col3	Col4
Data	Data	Data	Data
Data	Data	Data	Data
Data	Data	Data	Data

Document 1

```
{
 "prop1": data,
 "prop2": data,
 "prop3": data,
 "prop4": data
}
```

Document 2

```
{
 "prop1": data,
 "prop2": data,
 "prop3": data,
 "prop4": data
}
```

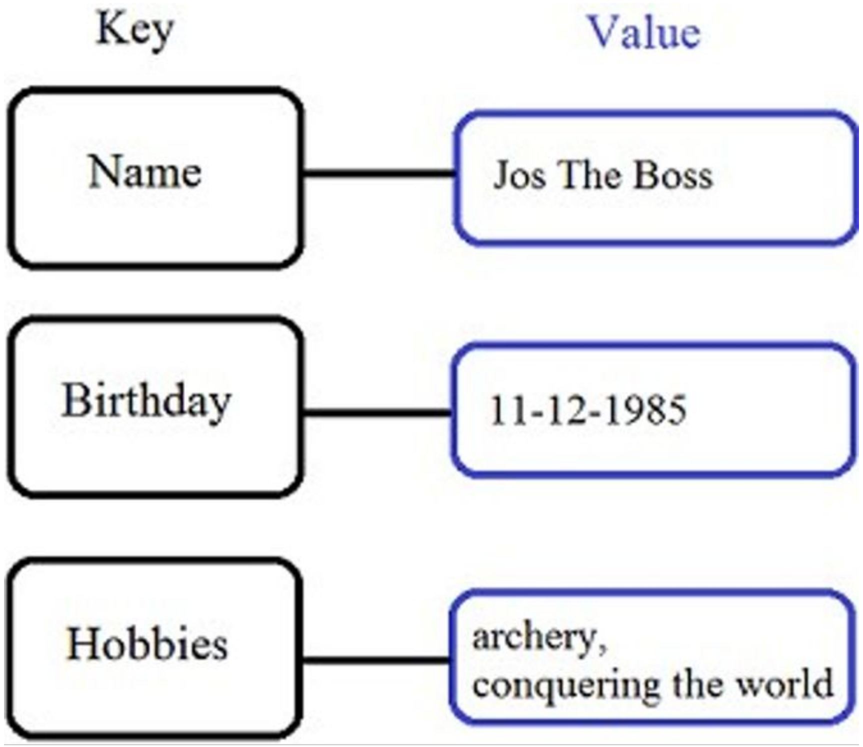
Document 3

```
{
 "prop1": data,
 "prop2": data,
 "prop3": data,
 "prop4": data
}
```

Document Based

- The database stores and retrieves documents, stored in the value part of the loosely structured sets of key-value pairs. e.g., XML, JSON etc.
- Self-describing, hierarchical tree data structures consisting of maps, collections and scalar values. Addressed in the db via a unique key
- The database offers an API or query language that retrieves documents based on their contents.
- Used for content management systems, blogging platforms, web analytics, real-time analytics, e-commerce applications.
- Avoided for systems that need complex transactions spanning multiple operations or queries against varying aggregate structures.
- Example: MongoDB, Couch DB, Lotus Notes, Orient DB, Raven DB

Key Value Pair Based

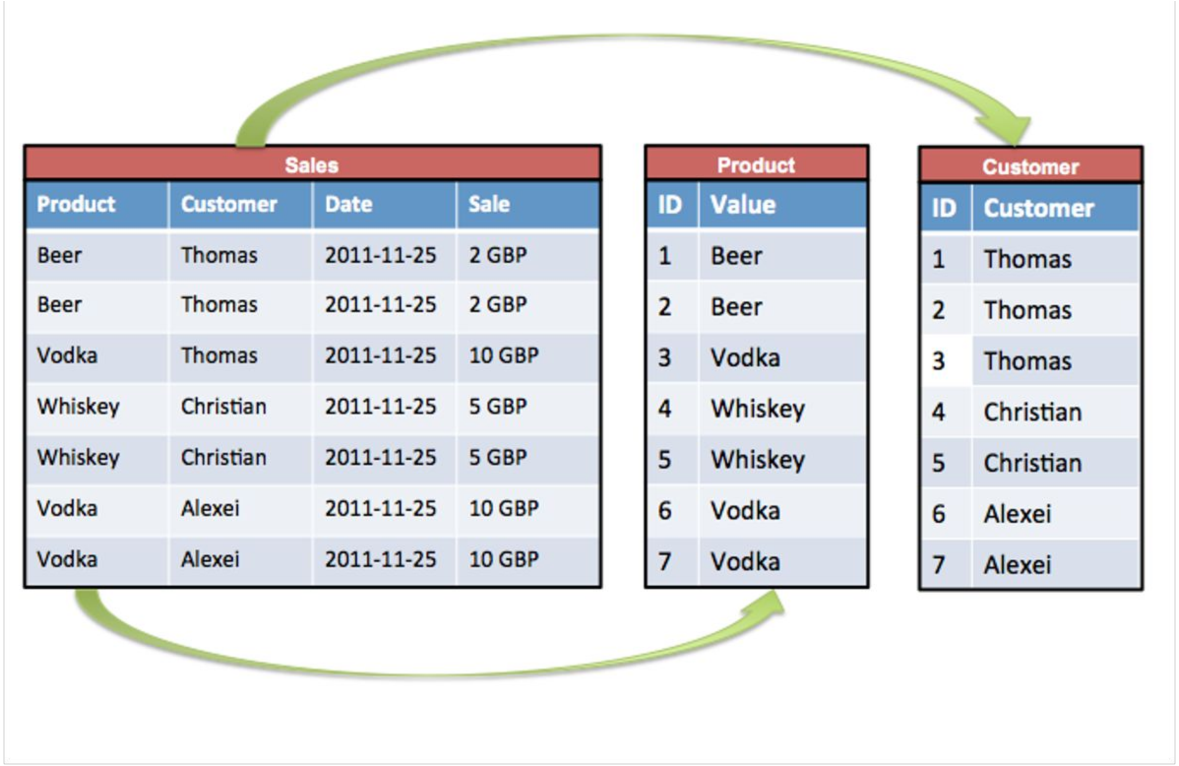


Key	Value
Name	Jos The Boss
Birthday	11-12-1985
Hobbies	archery, conquering the world

Key Value Pair Based

- Dictionaries contain a collection of records. Simplest of the four
- Store data as maps: Hashmaps or associative arrays. Provide efficient average running time algorithm for accessing data
- Records are stored and retrieved using a key that uniquely identifies the records and is used to quickly find the data within the database
- Used for storing session information, user profiles, preferences, shopping cart data etc.
- Avoided when we need to query data having relationships between entities.
- Example: Redis, CouchbaseDB, Oracle NoSQL database, Riak, Apache Cassandra, Amazon Dynamo, Voldemort

Column Based



Sales

Product	Customer	Date	Sale
Beer	Thomas	2011-11-25	2 GBP
Beer	Thomas	2011-11-25	2 GBP
Vodka	Thomas	2011-11-25	10 GBP
Whiskey	Christian	2011-11-25	5 GBP
Whiskey	Christian	2011-11-25	5 GBP
Vodka	Alexei	2011-11-25	10 GBP
Vodka	Alexei	2011-11-25	10 GBP

Product

ID	Value
1	Beer
2	Beer
3	Vodka
4	Whiskey
5	Whiskey
6	Vodka
7	Vodka

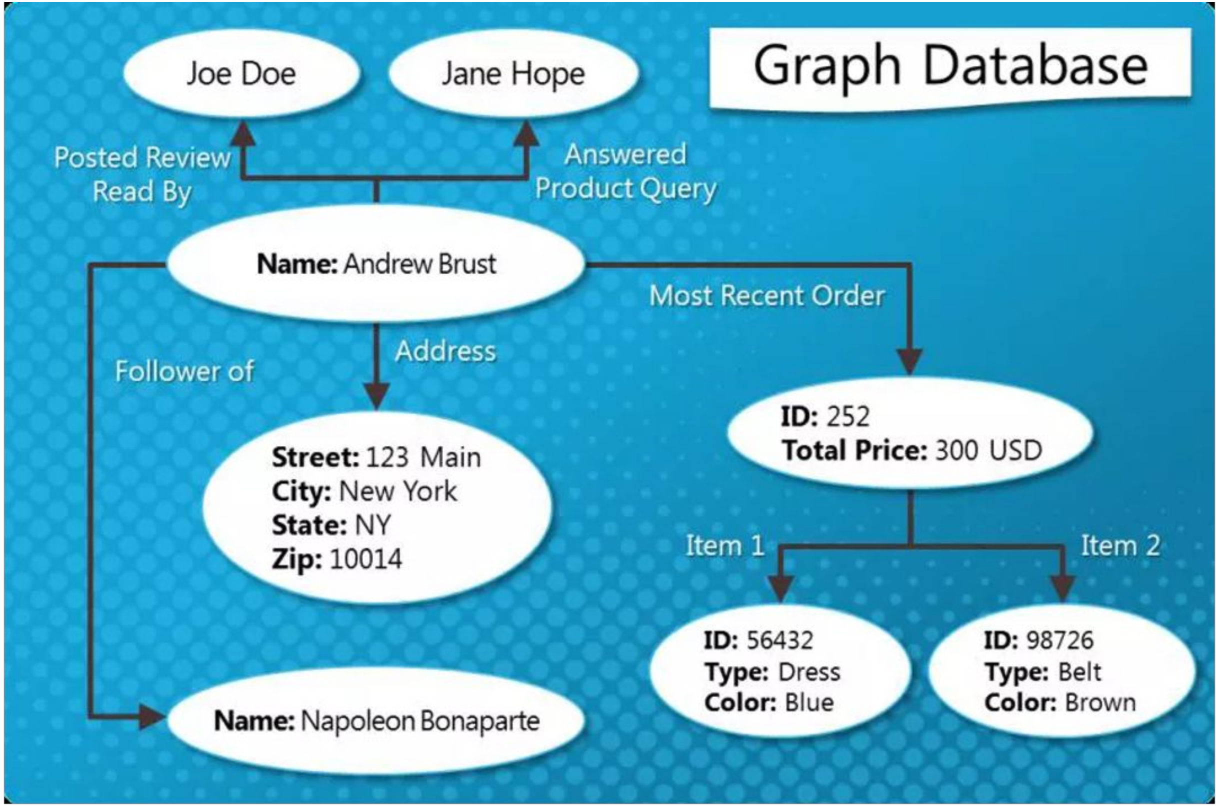
Customer

ID	Customer
1	Thomas
2	Thomas
3	Thomas
4	Christian
5	Christian
6	Alexei
7	Alexei

Column Based

- Data is efficiently stored in a column-oriented way.
- Columns are grouped in column-families. Data isn’t stored in a single table but in column families.
- Identified by “row-key”. Ordered and sorted based on row-key
- Used for content management systems, blogging platforms, log aggregation
- Avoided for systems that are in early development, changing query patterns.
- Example: Hbase, Cassandra, Hypertable, Amazon DynamoDB, Google’s Bigtable

Graph Based


Graph Database
Joe Doe
Jane Hope
Posted Review Read By
Answered Product Query
Name: Andrew Brust
Most Recent Order
Address
Follower of
ID: 252
Total Price: 300 USD
Street: 123 Main
City: New York
State: NY
Zip: 10014
Item 1
Item 2
ID: 56432
Type: Dress
Color: Blue
ID: 98726
Type: Belt
Color: Brown
Name: Napoleon Bonaparte

Graph Based

- Graph-oriented: Stored as an edge, a node or an attribute. Both nodes and edges can be labelled. Labels can used to narrow searches.
- Store entities and relationships between these entities as nodes and edges of a graph respectively.
- Traversing the relationships is very fast as relationship between nodes is not calculated at query time but is actually persisted as a relationship.
- Used for connected data, such as social networks, spatial data, routing information for goods and supply.
- Example: Neo4J, Infinite Graph, Orient DB, Flock DB

Apps and Use Cases

- Apple has reported they have a NoSQL database of over 75,000 servers, the largest known NoSQL database.
- During Amazon Prime Day in 2019, Their NoSQL database peaked at 45 million requests per second.

- ➔ Cloud providers, because scalability
  - ◆ DynamoDB, BigTable, CosmosDB
- ➔ Self-hosted
  - ◆ Cassandra, Scylla, CouchDB, MongoDB

Questions?

Thank you!